

Convex envelopes for fixed rank approximation

Fredrik Andersson ^{*}, Marcus Carlsson [†], and Carl Olsson [‡]

Centre for Mathematical Sciences, Lund University
Box 118, SE-22100, Lund, Sweden

Abstract

A convex envelope for the problem of finding the best approximation to a given matrix with a prescribed rank is constructed. This convex envelope allows the usage of traditional optimization techniques when additional constraints are added to the finite rank approximation problem. Expression for the dependence of the convex envelope on the singular values of the given matrix is derived and global minimization properties are derived. The corresponding proximity operator is also studied.

1 Introduction

Let $\mathbb{M}_{m,n}$ denote the Hilbert space of complex $m \times n$ -matrices equipped with the Frobenius (Hilbert-Schmidt) norm. The Eckart–Young–Schmidt theorem [4, 11] provides a solution to the classical problem of approximating a matrix by another matrix with a prescribed rank, i.e.,

$$\begin{aligned} \min \|A - F\|^2 \\ \text{subject to } \text{rank } A \leq K, \end{aligned} \tag{1.1}$$

by means of a singular value decomposition of F and keeping only the K largest singular vectors. However, if additional constraints are added then there will typically not be an explicit expression for the best approximation.

Let $g(A) = 0$ describe the additional constraints (for instance imposing a certain matrix structure on A), and consider

$$\begin{aligned} \min \|A - F\|^2 \\ \text{subject. to } \text{rank } A \leq K, \quad g(A) = 0. \end{aligned} \tag{1.2}$$

The problem (1.2) can be reformulated as

$$\begin{aligned} \min \mathcal{I}(A) = \mathcal{R}_K(A) + \|A - F\|^2 \\ \text{subject. to } g(A) = 0. \end{aligned} \tag{1.3}$$

^{*}fa@maths.lth.se

[†]mc@maths.lth.se

[‡]calle@maths.lth.se

where

$$\mathcal{R}_K(A) = \begin{cases} 0 & \text{rank } A \leq K, \\ \infty & \text{else.} \end{cases}$$

For instance, if g describes the condition that A is a Hankel matrix and F is the Hankel matrix generated by some vector f , then the minimization problem above is related to that of approximating f by K exponential functions [6]. This particular case of (1.3) was for instance studied in [1].

Standard (e.g. gradient based) optimization techniques do not work on (1.3) due to the highly discontinuous behavior of the rank function. A popular approach is to relax the optimization problem by replacing the rank constraint with a nuclear norm penalty, i.e. to consider the problem

$$\begin{aligned} & \mu_K \|A\|_* + \|A - F\|^2 \\ & \text{subject to } g(A) = 0. \end{aligned} \tag{1.4}$$

where $\|A\|_* = \sum_j \sigma_j(A)$, where the parameter μ_K is varied until the desired rank K is obtained.

In contrast to $\mathcal{R}_K(A)$ the nuclear norm $\|A\|_*$ is a convex function, and hence (1.4) is much easier to solve than (1.3). In fact, the nuclear norm is the convex envelope of the rank function (acting on matrices with operator norm ≤ 1) [5] which motivates the replacement of $\mathcal{R}_K(A)$ with $\mu_K \|A\|_*$ (for a suitable choice of μ_K).

However, the solutions obtained by solving this relaxed problem are often not good enough as approximations of the original problem. In fact the relaxation with replacing $\mathcal{R}_K(A)$ with $\mu_K \|A\|_*$ is not optimal even though the nuclear norm is the convex envelope of the rank function. This is because the contribution of the (convex) misfit term $\|A - F\|^2$ is not used. In [7, 8] it was suggested to incorporate the misfit term and work with the convex envelopes of

$$\mu \text{rank}(A) + \|A - F\|^2, \tag{1.5}$$

and

$$\mathcal{R}_K(A) + \|A - F\|^2, \tag{1.6}$$

respectively for the problem of low-rank and fixed rank approximations. The superior performance of using this relaxation approach in comparison to the nuclear norm approach was also verified by several examples in [7, 8]. For the convex envelope of (1.5) it turns out that there are simple explicit formulas acting on each of the singular values of F individually. This is not the case for the convex envelope of (1.6). Nevertheless, in [7, 8] an efficient optimization algorithm is presented that acts only on the singular values of F .

In this paper we present explicit expressions for the convex envelope of (1.6) in terms of the singular values $(\alpha_j)_{j=1}^{\min(m,n)}$ of A , as well a detailed information about global minimizers. More precisely, in Theorem 1 we show that the convex envelope of (1.6) is given by

$$\frac{1}{k_*} \left(\sum_{j>K-k_*} \alpha_j \right)^2 - \sum_{j>K-k_*} \alpha_j^2 + \|A - F\|^2. \tag{1.7}$$

where k_* is a particular value between 1 and K . To determine this value one uses Lemma 1 (Section 2). The second main result of this note is Theorem 2, where the global minimizers of (1.7) are found. In case the K -th singular value of F (denoted ϕ_K) has multiplicity one, then the minimizer of (1.7) is unique and coincides with that of (1.6), given by the Eckart-Young-Schmidt theorem. If ϕ_K has multiplicity M and is constant between sub-indices $J \leq K \leq L$, it turns out that the singular

values α_j of global minimizers A , in the range $J \leq j \leq L$ lie on a certain simplex in \mathbb{R}^M . We refer to Section 3 and (3.5) for further details.

In Section 4 we investigate the properties of the proximal operator

$$A \mapsto \operatorname{argmin}_A \mathcal{R}_K(A) + (1 + \rho)\|A - F\|^2, \quad \rho > 0.$$

In particular we show that it is a contraction with respect to the Frobenius norm and show that the proximal operator coincides with the solution of (1.1) whenever F has a sufficient gap between the K :th and $K + 1$:th singular value (see (4.4)).

2 Fenchel conjugates and the convex envelope

The Fenchel conjugate, also called the Legendre transform [10, Section 26], of a function f is defined by

$$f^*(B) = \operatorname{argmax}_A \langle A, B \rangle - f(A).$$

Note that for any function $f : \mathbb{M} \rightarrow \mathbb{R}$ that only depends on the singular values, we have that the maximum of $\langle A, B \rangle - f(A)$ with respect to A is achieved for a matrix A with the same Schmidt-vectors (singular vectors) as B , by von-Neumann's inequality [9]. More precisely, denote the singular values of A, B by α, β and denote the singular value decomposition by $A = U_A \Sigma_\alpha V_A^*$, where Σ_α is a diagonal matrix of length $N = \min(m, n)$. We then have:

Proposition 1. *For any $A, B \in \mathbb{M}_{m,n}$ we have $\langle A, B \rangle \leq \sum_{j=1}^N \alpha_j \beta_j$ with equality if and only if the singular vectors can be chosen such that $U_A = U_B$ and $V_A = V_B$.*

See [3] for a discussion regarding the proof and the original formulation of von Neumann. To simplify the presentation, in what follows we shall occasionally write $\mathcal{R}_K(\alpha)$ in place of $\mathcal{R}_K(\Sigma_\alpha)$ when it is suitable.

Proposition 2. *Let \mathcal{I} be as defined by (1.3). For its Fenchel conjugate it then holds that*

$$\mathcal{I}^*(B) = \sum_{j=1}^K (\sigma_j(F + B/2))^2 - \|F\|^2.$$

Proof. We abbreviate $\sigma_j(F + \frac{B}{2}) = \gamma_j$. Then

$$\mathcal{I}^*(B) = \sup_A \langle A, B \rangle - \mathcal{R}_K(A) - \|A - F\|^2 = \sup_\alpha -\mathcal{R}_K(\alpha) - \sum_{j=1}^N (\alpha_j - \gamma_j)^2 + \sum_{j=1}^N \gamma_j^2 - \|F\|^2.$$

It is clear that it is optimal to choose $\alpha_j = \gamma_j$ for $1 \leq j \leq K$ and $\alpha_j = 0$ otherwise. Hence,

$$\mathcal{I}^*(B) = - \sum_{j=K+1}^N \gamma_j^2 + \sum_{j=1}^N \gamma_j^2 - \|F\|^2. \quad \square$$

\square

For the computation of \mathcal{I}^{**} some auxiliary results are needed.

Lemma 1. *Let $(r_j)_{j=1}^K$ be an increasing sequence, $c \geq 0$, and set*

$$s_n = \frac{c + \sum_{j=1}^n r_j}{n}.$$

There exists a $1 \leq k_ \leq K$ such that this is a decreasing sequence of $1 \leq n \leq k_*$ and a strictly increasing sequence of $k_* \leq n \leq K$. Defining $r_{K+1} = \infty$ we have that k_* is the smallest value of n such that $s_n < r_{n+1}$ holds, and the largest value of n such that $r_n \leq s_n$ holds. In particular, it is the unique value satisfying*

$$r_{k_*} \leq s_{k_*} < r_{k_*+1}. \quad (2.1)$$

Proof. Let k_* be the first (i.e. smallest) value of n such that

$$s_n < r_{n+1}. \quad (2.2)$$

Note that

$$s_{n+1} = \frac{1}{n+1} r_{n+1} + \frac{n}{n+1} s_n$$

which is an weighted average, so s_{n+1} lies between r_{n+1} and s_n . As long as (2.2) fails we thus have

$$r_{n+1} \leq s_{n+1} \leq s_n, \quad (2.3)$$

and reversely

$$r_{n+1} > s_{n+1} > s_n, \quad (2.4)$$

when (2.2) holds. By (2.4) and the fact that $r_{n+2} \geq r_{n+1}$, we see that once (2.2) is fulfilled for some n , it is fulfilled for all subsequent n . This combined with the inequalities (2.3), (2.4) proves the first part of the statement.

For the remaining statements, note that we have already chosen k_* as the smallest value for which $s_n < r_{n+1}$ holds. Moreover, before this point (i.e. $n \leq k_*$) we do have $r_n \leq s_n$ by (2.3) (or, in case, $n = 1$, by the definition of s_1) and after this point (i.e. $n \geq k_*$) we do not have it, by (2.4) and the fact that this holds for all $n \geq k_*$, as noted earlier. Hence k_* is the largest value of n such that $r_n \leq s_n$ holds. The inequality (2.1) and its uniqueness now immediately follows. \square \square

For easier reference, we reformulate the above lemma in the setting it will be used.

Lemma 2. *Let $\beta \in \mathbb{R}^N$ be decreasing, $K < N$ fixed and set*

$$\omega_k = \frac{\sum_{j>K-k} \beta_j}{k}.$$

There exists a $1 \leq k_ \leq K$ such that ω is a decreasing sequence of $1 \leq n \leq k_*$ and a strictly increasing sequence of $k_* \leq n \leq K$. Defining $\beta_0 = \infty$ we have that k_* is the smallest value of k such that $\omega_k < \beta_{K-k}$ holds, and the largest value of k such that $\beta_{K-k+1} \leq \omega_k$ holds. In particular, it is the unique value satisfying*

$$\beta_{K-k_*+1} \leq \omega_{k_*} < \beta_{K-k_*}. \quad (2.5)$$

Proof. Apply Lemma 1 with $c = \sum_{j>K} \beta_j$ and $r_j = \beta_{K+1-j}$. \square

We are now ready to address (1.7).

Theorem 1. *The convex envelope of $\mathcal{R}_K(A) + \|A - F\|^2$ is*

$$\mathcal{I}^{**}(A) = \frac{1}{k_*} \left(\sum_{j>K-k_*} \alpha_j \right)^2 - \sum_{j>K-k_*} \alpha_j^2 + \|A - F\|^2$$

where $k_* = k_*(\alpha)$ ($1 \leq k_* \leq K$) is obtained by applying Lemma 2 with $\beta = \alpha$.

Note that $\frac{1}{k_*} \left(\sum_{j>K-k_*} \alpha_j \right)^2 = k_* \omega_{k_*}^2$ in the terminology of Lemma 2. Also note that $\mathcal{I}(A) \geq \|A - F\|^2$ and $\|A - F\|^2$ is convex in A . Since $\mathcal{I}^{**}(A)$ is the largest convex lower bound on $\mathcal{I}(A)$ we therefore have $\mathcal{I}^{**}(A) \geq \|A - F\|^2$ which shows that $\frac{1}{k_*} \left(\sum_{j>K-k_*} \alpha_j \right)^2 - \sum_{j>K-k_*} \alpha_j^2 \geq 0$.

Proof. We again employ the notation $\sigma_j \left(F + \frac{B}{2} \right) = \gamma_j$. For the bi-conjugate it then holds that

$$\begin{aligned} \mathcal{I}^{**}(A) &= \sup_B \langle A, B \rangle - \sum_{j=1}^K \gamma_j^2 + \|F\|^2 = \sup_B 2 \langle A, F + \frac{B}{2} \rangle - \sum_{j=1}^K \gamma_j^2 + \|A - F\|^2 - \|A\|^2 \\ &= \sup_{\gamma} 2 \sum_{j=1}^N \alpha_j \gamma_j - \sum_{j=1}^K \gamma_j^2 + \|A - F\|^2 - \|A\|^2 \\ &= \sup_{\gamma} 2 \sum_{j=K+1}^N \alpha_j \gamma_j - \left(\sum_{j=1}^K (\gamma_j - \alpha_j)^2 - \alpha_j^2 \right) + \|A - F\|^2 - \|A\|^2 \end{aligned}$$

where the middle identity follows by von-Neumann's trace inequality. We now hold γ_K fixed and consider the supremum over the remaining variables. Given $0 \leq k \leq K$ consider

$$\gamma_K \in [\alpha_{K-k+1}, \alpha_{K-k}] \quad (2.6)$$

(where as before $\alpha_0 = \infty$). It is not hard to see that the maximal value over the other variables is achieved by setting $\gamma_j = \gamma_K$ for $j > K - k$, and $\gamma_j = \alpha_j$ for the remaining ones. This gives

$$\begin{aligned} &\sup_{\{\gamma_j, j \neq K\}} 2 \sum_{j=K+1}^N \alpha_j \gamma_j - \left(\sum_{j=1}^K (\gamma_j - \alpha_j)^2 - \alpha_j^2 \right) \\ &= \sup_{\{\gamma_j, j \neq K\}} 2 \sum_{j=K+1}^N \alpha_j \gamma_K - \left(\sum_{j=K-k+1}^K (\gamma_K - \alpha_j)^2 - \alpha_j^2 \right) + \left(\sum_{j=1}^{K-k} \alpha_j^2 \right) \\ &= 2\gamma_K \sum_{j=K-k+1}^N \alpha_j - k\gamma_K^2 + \left(\sum_{j=1}^{K-k} \alpha_j^2 \right) := f(\gamma_K), \end{aligned} \quad (2.7)$$

Since f by definition is defined as a partial supremum over a concave function, it follows that f itself is concave (see e.g. Section 3.2.5 in [2]). In particular, the different expressions valid in the different regimes (2.6) agree at overlapping endpoints. Also, the expression for $k = 0$ is valid in $[0, \alpha_K]$, and since this is linear non-decreasing, the supremum of f is attained in one of the other

intervals (possibly at α_K). We may thus assume that the supremum is attained in a (non-void) interval of the form

$$\gamma_K \in [\alpha_{K-k+1}, \alpha_{K-k}] \quad (2.8)$$

with $k \geq 1$. By fixing k and differentiating the expression for f in (2.7), we conclude that the maximum is obtained at the point

$$\omega_k = \frac{\sum_{j=K-k+1}^N \alpha_j}{k}$$

which then must lie in the interval (2.8). With $\beta = \alpha$, this constraint is precisely the inequalities (2.5), and hence appropriate k equals k_* given by Lemma 2 applied to α . Moreover, by (2.7) we then get

$$\begin{aligned} \sup_{\gamma} 2 \sum_{j=K+1}^N \alpha_j \gamma_j - \left(\sum_{j=1}^K (\gamma_j - \alpha_j)^2 - \alpha_j^2 \right) &= \sup_{\gamma_K} f(\gamma_K) = f(\omega_{k_*}) \\ &= 2\omega_{k_*} \sum_{j=K-k_*+1}^N \alpha_j - k_* \omega_{k_*}^2 + \left(\sum_{j=1}^{K-k_*} \alpha_j^2 \right) = k_* \omega_{k_*}^2 + \left(\sum_{j=1}^{K-k_*} \alpha_j^2 \right). \end{aligned}$$

Returning to the initial calculation we thus see that

$$\mathcal{I}^{**}(A) = k_* \omega_{k_*}^2 + \left(\sum_{j=1}^{K-k_*} \alpha_j^2 \right) + \|A - F\|^2 - \|A\|^2$$

which proves the theorem since $\|A\|^2 = \sum_{j=1}^N \alpha_j^2$. \square

3 Global minimizers

We now consider global minimizers of \mathcal{I} and \mathcal{I}^{**} . Given a sequence $(\phi_n)_{n=1}^N$ we recall that Σ_ϕ denotes the corresponding diagonal matrix. We introduce the notation $\tilde{\phi}$ for the sequence ϕ truncated at K , i.e.

$$\tilde{\phi}_j = \begin{cases} \phi_j & \text{if } 1 \leq j \leq K, \\ 0 & \text{otherwise.} \end{cases} \quad (3.1)$$

Since K is a fixed number which is clear from the context, we will usually abbreviate $\tilde{\phi}$ by ϕ . Recall the Eckart-Young-Schmidt theorem, which can be rephrased as follows;

The solutions to $\arg\min_A \mathcal{I}(A)$ are all matrices of the form $A_ = U \Sigma_\phi^A V^*$, where $U \Sigma_\phi V^*$ is any singular value decomposition of F . A_* is unique if and only if the singular value ϕ_K has multiplicity one.*

Obviously, a global minimizer of \mathcal{I} is a global minimizer of \mathcal{I}^{**} , but the converse need not be true. It is not hard to see that, in case ϕ_K has multiplicity one, the minimizer of \mathcal{I} is also the (unique) minimizer of \mathcal{I}^{**} . The general situation is more complicated. Given integers m and M with $m \leq M$, denote by $\Omega_{M,m}$ the simplex in \mathbb{R}^M given by the hyperplane

$$\sum_{j=1}^M x_j = m \quad (3.2)$$

and the inequalities

$$1 \geq x_1 \geq x_2 \geq \dots \geq x_M \geq 0. \quad (3.3)$$

Theorem 2. *Let $K \in \mathbb{N}$ be given, let F be a fixed matrix and let ϕ be its singular values. Let ϕ_J (respectively ϕ_L) be the first (respectively last) singular value that equals ϕ_K , and set $M = L + 1 - J$ (that is, the multiplicity of ϕ_K). Finally set $m = K + 1 - J$, (that is, the multiplicity of $\tilde{\phi}_K$).*

*The global minimum of \mathcal{I} and \mathcal{I}^{**} both equal $\sum_{j>K} \phi_j^2$ and the solutions to*

$$\operatorname{argmin}_A \mathcal{I}^{**}(A) \quad (3.4)$$

are all matrices of the form $A_ = U \Sigma_\alpha V^*$, where $U \Sigma_\phi V^*$ is any singular value decomposition of F , and α is a decreasing sequence satisfying:*

$$\begin{cases} \alpha_j = \phi_j, & 1 \leq j < J, \\ (\alpha_j)_{j=J}^L \in \phi_K \Omega_{M,m}, \\ \alpha_j = 0, & j > L. \end{cases} \quad (3.5)$$

In particular, A_ is unique if and only if ϕ_K has multiplicity one. Also, the maximal rank of such an A_* is L and the minimal rank is J .*

Proof. The fact that the minimum value of \mathcal{I} and \mathcal{I}^{**} coincide follows immediately since \mathcal{I}^{**} is the convex envelope of \mathcal{I} , and the fact that this value is $\sum_{j>K} \phi_j^2$ follows by the Eckart-Young-Schmidt theorem.

Let A be a solution to (3.4). By Proposition 1 it then follows that we can choose matrices U and V such that $A = U \Sigma_\alpha V^*$ and $F = U \Sigma_\phi V^*$ are singular value decompositions of A and F respectively. Set $\tilde{F} = U \Sigma_{\tilde{\phi}} V^*$. Note that \tilde{F} also is a minimizer of (3.4), which follows by the first sentence of the proof and the fact that $\mathcal{I}(\tilde{F}) = \sum_{j>K} \phi_j^2$. Since \mathcal{I}^{**} is the convex envelope of \mathcal{I} , it follows that all matrices

$$A(t) = \tilde{F} + t(A - \tilde{F}), \quad 0 \leq t \leq 1,$$

are solutions of (3.4). Set

$$\epsilon = \alpha - \tilde{\phi}, \quad (3.6)$$

where α are the singular values of A and note that $A(t) = U \Sigma_{\tilde{\phi} + t\epsilon} V^*$. Since α is a decreasing non-negative sequence, we also get certain restrictions on ϵ such as $(\epsilon_j)_{j=J}^K$ being decreasing and $(\epsilon_j)_{j>K}$ being decreasing and non-negative.

We now compute $\mathcal{I}^{**}(A(t))$ according to Theorem 1 for some fixed value of t . Set

$$\alpha(t) = \tilde{\phi} + t\epsilon$$

and note that the condition for choosing $k = k_*(\alpha(t))$ is

$$\alpha_{K+1-k}(t) \leq \frac{\sum_{j=K+1-k}^N \alpha_j(t)}{k} < \alpha_{K-k}(t),$$

(where we abbreviate $\sum_{j=K+1-k}^N$ by \sum_{K+1-k}^N for simpler reading). For small values of t and $k \leq m$, all numbers above are close to ϕ_K , except $\alpha_{K-k}(t)$ for the value $k = m$, in which case $\alpha_{K-k}(t) \approx \phi_{J-1}$ which by choice of J is strictly larger than ϕ_K . It follows by Lemma 2 that

$k_*(\alpha(t)) \leq m$ for values of t near 0. We only consider such values in what follows. By Lemma 2 we also have that $k_*(\alpha(t))$ equals the largest value for which

$$\alpha_{K+1-k}(t) \leq \frac{\sum_{K+1-k}^N \alpha_j(t)}{k} \quad (3.7)$$

holds. Since $\alpha_j(t) = \phi_K + t\epsilon_j$ for $J \leq j \leq K$, it easily follows that (3.7) is equivalent with

$$\epsilon_{K+1-k} \leq \frac{\sum_{K+1-k}^N \epsilon_j}{k}, \quad (3.8)$$

which is independent of t , and hence so is $k_*(\alpha(t))$. In the remainder we simply write k_* . It follows that

$$\begin{aligned} \mathcal{I}^{**}(A(t)) &= k_* \left(\phi_K + t \frac{\sum_{K+1-k_*}^N \epsilon_j}{k_*} \right)^2 - \sum_{K+1-k_*}^K (\phi_K + t\epsilon_j)^2 - \sum_{K+1}^N (t\epsilon_j)^2 + \|A(t) - F\|^2 \\ &= k_* \left(\phi_K + t \frac{\sum_{K+1-k_*}^N \epsilon_j}{k_*} \right)^2 - \sum_{K+1-k_*}^K (\phi_K + t\epsilon_j)^2 - \sum_{K+1}^N (t\epsilon_j)^2 + \sum_1^K (t\epsilon_j)^2 + \sum_{K+1}^N (t\epsilon_j - \phi_j)^2, \end{aligned}$$

which looks like a second degree polynomial in t (with constant term $\sum_{j>K} \phi_j^2$ as it should). However, note that this polynomial is in fact constant by our assumption on $A(t)$, and hence the first and second coefficient are zero. The coefficient of the linear term is

$$2\phi_K \sum_{K+1-k_*}^N \epsilon_j - 2\phi_K \sum_{K+1-k_*}^K \epsilon_j - 2 \sum_{K+1}^N \phi_j \epsilon_j = 2 \sum_{K+1}^N (\phi_K - \phi_j) \epsilon_j.$$

Note that $\phi_K - \phi_j = 0$ for $K < j \leq L$, that $\phi_K - \phi_j > 0$ for $j > L$ and that ϵ_j is non-negative in this range. We conclude that

$$\epsilon_j = 0, \quad j > L. \quad (3.9)$$

The coefficient of the quadratic term is

$$k_* \left(\frac{\sum_{K+1-k_*}^N \epsilon_j}{k_*} \right)^2 - \sum_{K+1-k_*}^N \epsilon_j^2 + \sum_1^K \epsilon_j^2 = k_* \left(\frac{\sum_{K+1-k_*}^L \epsilon_j}{k_*} \right)^2 + \sum_1^{K-k_*} \epsilon_j^2$$

where we have used (3.9). It clearly follows that

$$\sum_{K+1-k_*}^L \epsilon_j = 0, \quad \epsilon_j = 0, \quad 1 \leq j \leq K - k_*. \quad (3.10)$$

If k_* is not maximal, i.e. equal to m , then ϵ_J is among the ϵ_j 's in (3.10), which forces $(\epsilon_j)_J^K$ to be non-positive since this is a decreasing sequence, as noted following (3.6). If k_* is maximal then $K+1-k_* = J$ and (3.8), (3.10) implies that

$$\epsilon_J \leq \frac{\sum_J^L \epsilon_j}{k_*} = 0,$$

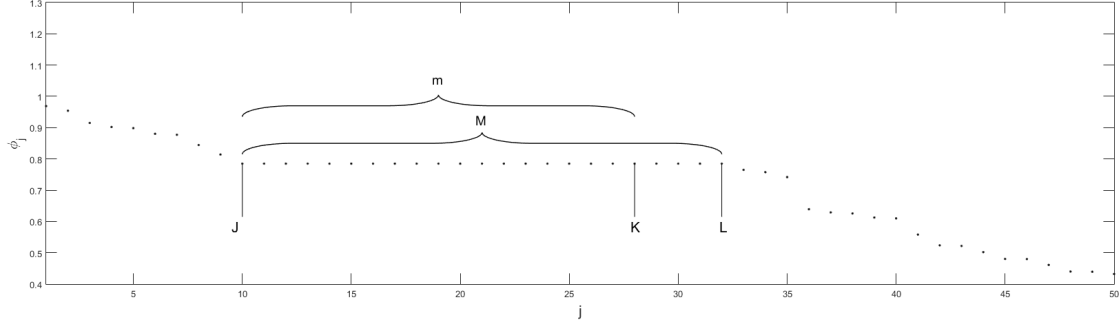


Figure 1: Illustration of the notation used in Theorem 2

from which the same conclusion follows. Summing up, the above, equations (3.9), (3.10), and the remarks following (3.6), together imply that $(\epsilon_j)_1^N$ is zero except possibly in the interval $\{J, \dots, K\}$, where it is decreasing non-positive, and the interval $\{K + 1, \dots, L\}$, where it is decreasing non-negative.

The entire analysis has been valid for “small” t , since at the outset we could not prove that $k_* \leq m$ for all values of t . By now we know that $\alpha_{J-1}(t) = \phi_{J-1}$ and that $\alpha_J(t)$ is decreasing, by which it follows that the whole previous analysis is valid in the entire range $t \in (0, 1]$. The top affirmation of (3.5) now follows by (3.10), and the bottom follows by (3.9). It remains to prove the middle. Clearly $(\alpha_j(0))_j^L$ lies in the hyperplane (3.2) since the first m values equal ϕ_K and the remaining ones are 0. From (3.10) it is clear that $\sum_J^L \epsilon_j = 0$, so we stay in this hyperplane as t varies. Concerning (3.3), that $(\alpha_j(t))_j^L$ has to be increasing is immediate by construction, and $\alpha_J(t) \leq 1 \cdot \phi_K$ follows as $\epsilon_J \leq 0$, as noted earlier. We conclude that if A is a solution to (3.4), then (3.5) is satisfied.

Conversely, if it has this form then the calculations of the proof clearly shows that $\mathcal{I}^{**}(A(t))$ is a constant polynomial equal to $\sum_{j>K} \phi_j^2$. Finally, the uniqueness statement and the rank statements are immediate. \square

4 The proximal operator

Theorem 3. *Let $F = U_F \Sigma_\phi V_F^*$ be given. The solution of*

$$\operatorname{argmin}_A \mathcal{I}^{**}(A) + \rho \|A - F\|^2, \quad (4.1)$$

is of the form $A = U_F \Sigma_\alpha V_F^$ where α has the following structure; there exists natural numbers $k_1 \leq K \leq k_2$ and real number $s > \phi_{k_2}$ such that*

$$\alpha_j = \begin{cases} \phi_j, & j < k_1 \\ \phi_j - \frac{s - \phi_j}{\rho}, & k_1 \leq j \leq k_2 \\ 0, & j > k_2 \end{cases} \quad (4.2)$$

In particular, α is a decreasing sequence and $\alpha \leq \phi$. In other words, the proximal operator is a contraction.

The theorem can be deduced by working directly with the expression for \mathcal{I}^{**} , but it turns out that it is easier to follow the approach in [7] which is based on the minimax theorem and an analysis of the simpler functional \mathcal{I}^* . We give more concrete information about how to find s , k_1 and k_2 in a separate proposition after the proof.

Proof. The proof partially follows the approach in [7], Section 3.1. Using Proposition 2 and some algebraic simplifications, (4.1) can be rewritten

$$\begin{aligned} & \operatorname{argmin}_A \max_B \langle A, B \rangle - \mathcal{I}^*(B) + \rho \|A - F\|^2 = \\ & \operatorname{argmin}_A \max_B \langle A, B \rangle - \sum_{j=1}^K \left(\sigma_j \left(F + \frac{B}{2} \right) \right)^2 + \|F\|^2 + \rho \|A - F\|^2 \\ & \operatorname{argmin}_A \max_Z \rho \left\| A - \frac{(1+\rho)F - Z}{\rho} \right\|^2 - \frac{1}{\rho} \|Z - (1+\rho)F\|^2 + (1+\rho) \|F\|^2 - \sum_{j=1}^K (\sigma_j(Z))^2. \end{aligned}$$

Switching the order of max and min gives the relation $A = ((1+\rho)F - Z)/\rho$, and this in turn yields that the maximization over Z takes the form

$$\operatorname{argmax}_Z -\frac{1}{\rho} \|Z - (1+\rho)F\|^2 - \sum_{j=1}^K \zeta_j^2, \quad \text{where } \zeta_j = \sigma_j(Z).$$

By Proposition it follows that the appropriate Z shares singular vectors with F , so the problem reduces to that of minimizing

$$\operatorname{argmin}_{\zeta} \sum_{j=1}^N (\zeta_j - (1+\rho)\phi_j)^2 + \rho \sum_{j=1}^K \zeta_j^2 = \operatorname{argmin}_{\zeta} (1+\rho) \sum_{j=1}^K (\zeta_j - \phi_j)^2 + \sum_{j=K+1}^N (\zeta_j - (1+\rho)\phi_j)^2.$$

The unconstrained minimization (i.e. ignoring that the singular values need to be decreasing) of this is $\zeta_j = \phi_j$ for $j \leq K$ and $\zeta_j = (1+\rho)\phi_j$ for $j > K$. It is proven in the appendix of [7] that the constrained minimization has the solution

$$\zeta_j = \begin{cases} \max(\phi_j, s), & j \leq K \\ \min((1+\rho)\phi_j, s), & j > K \end{cases} \quad (4.3)$$

where s is a parameter between ϕ_K and $(1+\rho)\phi_{K+1}$. The appropriate value of s is easily found by inserting this into the expression above. Let k_1 resp. k_2 be the first resp. last index where s shows up in ζ . Formula (4.2) is now an easy consequence of (4.3). \square

Proposition 3. *The appropriate value of s is found by minimizing*

$$\sum_{j=1}^K (\max(\phi_j, s) - \phi_j)^2 + \sum_{j=K+1}^N \left(\min(\phi_j, \frac{s}{1+\rho}) - \phi_j \right)^2.$$

in the interval $[\phi_K, (1+\rho)\phi_{K+1}]$. Given such an s , k_1 is the smallest index ϕ with $\phi_{k_1} < s$ and k_2 last index with $\phi_{k_2} > \frac{s}{1+\rho}$.

Note in particular that the proximal operator (given by Theorem 3) reduce to (3.1) if

$$\phi_K \geq (1+\rho)\phi_{K+1}. \quad (4.4)$$

5 Conclusions

We have analyzed and derived expressions for how to compute the convex envelope corresponding to the problem of finding the best approximation to a given matrix with a prescribed rank. These expressions work directly on the singular values.

6 Acknowledgements

This research is partially supported by the Swedish Research Council, grants no. 2011-5589, 2012-4213 and 2015-03780; and the Crafoord Foundation.

References

- [1] Fredrik Andersson, Marcus Carlsson, Jean-Yves Tournieret, and Herwig Wendt. A new frequency estimation method for equally and unequally spaced data. *Signal Processing, IEEE Transactions on*, 62(21):5761–5774, 2014.
- [2] S. Boyd and L. Vandenberghe. *Convex Optimization*. Cambridge University Press, 2004.
- [3] Eduardo Marques de Sá. Exposed faces and duality for symmetric and unitarily invariant norms. *Linear Algebra and its Applications*, 197:429–450, 1994.
- [4] Carl Eckart and Gale Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, 1936.
- [5] Maryam Fazel. *Matrix rank minimization with applications*. PhD thesis, PhD thesis, Stanford University, 2002.
- [6] Leopold Kronecker. *Zur Theorie der Elimination einer Variablen aus zwei algebraischen Gleichungen*. Königliche Akad. der Wissenschaften, 1881.
- [7] Viktor Larsson and Carl Olsson. Convex envelopes for low rank approximation. In *Energy Minimization Methods in Computer Vision and Pattern Recognition*, pages 1–14. Springer, 2015.
- [8] Viktor Larsson, Carl Olsson, Erik Bylow, and Fredrik Kahl. Rank minimization with structured data patterns. In *Computer Vision–ECCV 2014*, pages 250–265. Springer, 2014.
- [9] Leon Mirsky. A trace inequality of John von Neumann. *Monatshefte für Mathematik*, 79(4):303–306, 1975.
- [10] Ralph Tyrell Rockafellar. *Convex analysis*. Princeton university press, 2015.
- [11] Erhard Schmidt. Zur theorie der linearen und nichtlinearen integralgleichungen. III. Teil. *Mathematische Annalen*, 65(3):370–399, 1908.